Detection of transcription factor binding sites using Rényi entropy

Joan Maynou, Montserrat Vallverdú, Francesc Clarià, Alexandre Perera and Pere Caminal

***Abstract*—During the process of protein synthesis, transcription of DNA to messenger RNA starts with the binding of the transcription factors to the promoter. One of the issues on the prediction of transcription factor binding is that sequences corresponding to the binding present variability. In this manuscript a method for the detection of binding site is proposed, based on a parametric uncertainty measurement (Rényi entropy). This measurement is done through an estimation of the probability for each nucleotide avoiding any numerical representation of the nucleotides. We obtain values of the efficiency of the method as Receiver Operating Characteristic curves found on ABF1 and ROX1 binding sites in chromosome I and XVI of the organism *Saccharomyces cerevisiae*.**

I. Introduction

Molecular genetics establish that the information content in a gene can originate a protein using the processes of transcription and translation. Transcription is the process where the genetic information, initially as a deoxyribonucleic acid, DNA, results in ribonucleic acid messenger, mRNA. This process begins by means of the union of RNA polymerase enzyme and the transcription factors to the promoter, a nucleotide sequence which has the signal to start the transcription. Once the different molecules are joined with the promoter, the copy of a DNA strand to mRNA is triggered. In the eukaryotic cells, transcription and translation stages are not directly connected as the nuclear membrane physically separates the process. The mRNA obtained must be modified to leave the nucleus using the processes of 7-metilguanosina, polyadenylation and splicing. After the mRNA has been processed, it is translated into an amino acids sequence, process known as translation. These polypeptides or proteins form structural proteins and enzymes that control the metabolic processes in cells [1-2].

A single transcription factor shows binding among different sites, with different sequences. Due to this intrinsic variability it is difficult to establish a *consensus* sequence approach for binding detection [3]. Consequently, any detection method of binding sites within a DNA sequence, must consider the variability of these ones. This has originated several efforts of research, employing different methods to detect patterns in bio sequences: probabilistic, deterministic and numerical [4]. The probabilistic methods are characterized by the need of being trained, to be able then, to infer the discovery of patterns. Within this field the most representative models are based on Position Weight Matrix or PWM. These are based on the frequency of each symbol in a specific position of a training group. These models generally assume the independence between positions. On the other hand, models like Hidden Markov Models or HMM, and neuronal nets, assume the dependence between positions of a binding site also under probabilistic approaches [5], [6], [7], [8]. The deterministic methods are based on adjusting a specific sequence to concrete patterns [9], [10]. Some others use multivariate models and additional information that include the context [11]. Information theory has also been used in genetics to visualize the information of a sequence set [12], [13], [14], [15]. There is a first study about the characterization of a sequence set with parametric entropies [16], although there are no published results on the use of parametric entropies for building a detector.

In this manuscript we propose to detect binding sites of transcription factors using parametrical entropy. The method employs an aligned set of sequences with known binding and checks the total information change when the candidate sequence is included in the set.

II. Materials and Methods

A. Method

The proposed method starts with a matrix of aligned sequences with binding evidence. The transcription factor binding sites, TFBS, are detected in a candidate sequence by means of the training set information content in a specific position [16]. Any new candidate sequence added to the training matrix will cause a variation on the order or information of set of aligned sequences. For random sequences the disorder in the system will increase. For a true binding site, the candidate sequence is not expected to modify in a significant way the total information of the aligned sequence set.

The classical uncertainty or order measure is the Shannon Entropy. In this study, Rényi entropy is employed for this measurement which depends on the q parameter which is known as the order in the Rényi entropy. This parameter modulates the probability of occurrence of each symbol, emphasizing/suppressing this value as q decreases/increases. This measurement allows us to build a parametric detector

This work was supported by the Spanish Ministerio de Educación y Ciencia under the Ramón y Cajal Program and TEC2007-63637/TCM and the CIBER- Centro de Investigación Biomédcia en Red en Bioingeniería, Biomateriales y Nanomedicina.

J. Maynou, M. Vallverdú, A. Perera and P. Caminal are with Dep. ESAII, Centre for Biomedical Engineering Research, Technical University of Catalonia (UPC), Barcelona, Gargallo, 5, 08028 Barcelona, Spain. (e-mail: joan.maynou@upc.edu, Alexandre.Perera@upc.edu, Montserrat.Vallverdu@upc.edu and Pere.Caminal@upc.edu).

F. Clarià is Dep. d'Informàtica I Ingenyieria Industrial, Universitat de Lleida, Lleida, Spain (e-mail: Claria@eup.udll.es).

with variable sensibility thanks to the q parameter considered.

B. Database description

The algorithm requires a group of aligned nucleotide sequences with binding evidence. These sequences correspond to the organism *Saccharomyces cerevisiae*. *Saccharomyces cerevisiae* was the first eukaryotic organism with its genome completed. This organism contains around sixteen million of nucleotides distributed among sixteen chromosomes. We have considered the recognizers ROX1 and ABF1 (Table 1), located in the chromosomes I and XVI with chromosome length 948062 and 230208 nucleotides respectively. The dataset has been obtained from the data base TRANSFAC [17], http://www.gen-regulation.com/pub/databases.html, using for the extraction of DNA sequences, an R library for automatic sequence extraction from a transcription factor name. Finally, these sequences have been lined up by means of MUSCLE [18], to obtain the different nucleotides involved in each position.

TABLE I
SUMMARY OF THE RECOGNIZERS ANALYZED

ORGANISM	RECOGNIZER	BASES	ALIGNED SEQUENCES
S. cerevisiae	ROX1	12	20
S. cerevisiae	ABF1	37	22

C. Information content measures

The disorder in a system can be computed using the measure of Rényi entropy. The Rényi entropy [19] is considered a generalization of the Shannon entropy. With a random variable x with N possible states(x_1, x_2,…,x_N) and a probability for each state i, given by p_i, the Rényi entropy is defined as,

$$H_q = \frac{1}{1-q}\log_2\left(\sum_{i=1}^{N} p_i^q\right) \quad (1)$$

q is a positive real number different than 1 (also known as alpha parameter in [16]). Rényi entropy takes its maximum value when all possible states show equal probability $p_i = 1/N$. The opposite case Hq=0 occurs when a particular state as full probability. Rényi entropy converges to Shannon entropy when q tends to 1.

$$H_s = -\sum_{i=1}^{N} p_i \log(p_i) \quad (2)$$

The normalized redundancy R is defined as,

$$R = 1 - \frac{H}{H|_{\max}} \quad (3)$$

where the redundancy is normalized depending on the maximum entropy. R decreases with the increase of information, and therefore increases with the increase of order. For a group of aligned sequences, the measurement of the redundancy, Shannon or Rényi, in a specific PWM, gives information about the complexity of the nucleotides distribution in the conserved sequence.

D. TFBS detection

By means of a matrix of aligned sequences we perform a measurement of the order in the different positions of the binding sites using the Rényi entropy [20]. The values of redundancy for very variable positions are close to 0. On the other hand, for positions highly ordered redundancy has values close to the unity. Using this premise, the algorithm developed does a comparison between the redundancy profile and the redundancy profile of the matrix when the candidate sequence is added to the set. This comparison is done position by position using the multiplication between both profiles as shown in (4).

$$\Phi = R_{matrix} \cdot R_{[matrix+seq]} \quad (4)$$

$R_{[matrix+seq]}$ measurement determines the order of the system when we add the studied sequence. The redundancy profile is represented as an n-dimensional vector, where n is the total number of positions of the binding site. This creates a vector space of redundancies where each axis corresponds to a specific position of the binding site. For each vector, the 2-norm is calculated relative to the origin of coordinates (5).

$$\left\|\vec{\phi}\right\| = \sqrt{\phi_1^2 + \ldots + \phi_i^2 + \ldots + \phi_N^2} \quad (5)$$

The norm which corresponds to the redundancy vector of the training matrix is the maximum norm for this system. When we add the candidate sequence to the aligned matrix Φ will be equal or lower than the maximum norm. The closest the candidate sequence to the training set, the larger the value of Φ. This defines an index which allows for the discrimination between a random sequence and a sequence that belongs to a binding site.

The developed method, which is based in the criteria defined previously, is described next:

1. For each position within the training matrix, we estimate the probability of every nucleotide by means of the appearance frequency. We consider the missing values, using the expectancy of a random variable, Ω [*A, T, C, G*], with the probabilities of each nucleotide [21].
2. The redundancy profile is calculated from the PWM, correcting finite sample effects [22].
3. 1 and 2 are repeated, considering the training matrix with the new sequence added.
4. For each redundancy profile obtained from the studied sequences, we have to calculate the product between profiles and the norm which corresponds to them.
5. *p-value* is calculated regarding the null distribution of the norm. If $p<\alpha$, we consider that the sequence belongs to a binding site.

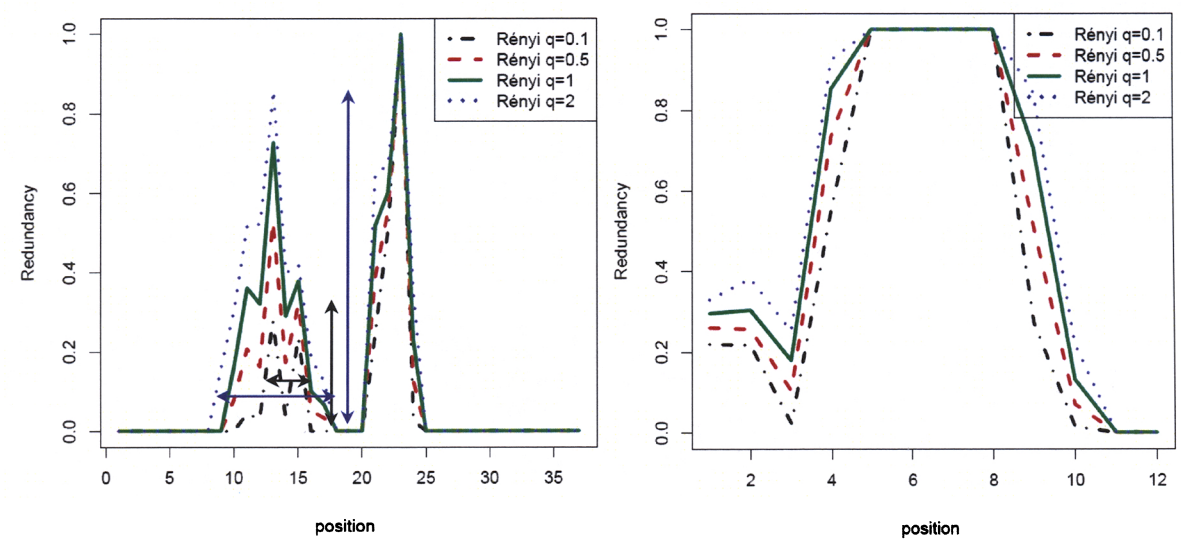

Fig.1. Left to right: redundancy for ABF1 and ROX1 binding sites.

III. Results

In the figure 1, we can observe the variability of each position of the ABF1 transcription factor by means of the correspondent redundancy profile for different q-values. The dependence of the entropic profiles with q is also shown.

The Rényi order modulates the amplitude and the number of positions that belong to a binding site, and then we can obtain the positions involved to the binding sites. As q increases, the noise in the redundancy signal increases too. With low q the redundancy signal also decreases. Therefore, the redundancy profile of the transcription factor depends on the Rényi order. An optimal q-value is suggested as a trade-off between the noise included in the redundancy signal and the attenuation of the same one.

The validation of the detector has been realized by means of "Leave one-out cross-validation ". Every individual sequence is used as a test sequence of training classifier with the rest of N-1 sequences. First results have been obtained with randomly generated candidate sequences. A null distribution for Φ is obtained by testing 1000 times a random sequence on the rest of N-1 sequences. That's made successively for each sequence within the training matrix. For the real chromosomes I and XVI from *S. cerevisiae*, the distribution is obtained by testing the rest of the N-1 remaining sequences on the chromosome, I and the XVI. The performance of the detection is shown as a Receiver Operating Characteristic (ROC) curve for different q on Figure 2. We observe that the number of TP, true positives, and FP, false positives, depend on the TFBS. In the same number of TP, the number of FP is bigger in the case ABF1 than in ROC1. This is because that the number of positions involved in the binding site is bigger in ROC1.

In the table 2, we observe that the detector has a different behaviour depending on the q value. The best learning system will be that one which produces a bigger area under the convex surface, AUC. If q decreases, the number of positions of the transcription factor that we consider decreases, but the number of TP and FP increases. On the other hand, if q increases, increases the number positions of the transcription factor and also the FP, but the TP decreases. Therefore, the Rényi order q does depend on the TFBS characteristics and should be adjusted for each training sequence set (e.g. by means of cross-validation).

TABLE 2
AUC FOR ABF1 AND ROX1 FOR RANDOM SEQUENCES AND *S. CEREVISIAE* CR. I AND XVI

	RANDOM		REAL	
	ABF1	ROX1	ABF1	ROX1
q=0.1	0.9843	0.9992	0.9251	0.9836
q=0.5	0.9882	0.9992	0.9315	0.9833
q=1.0	0.9892	0.9988	0.9238	0.9807
q=2.0	0.9895	0.9989	0.8917	0.9773

Generally, low q values will depress the Redundancy profile, turning Φ more selective, whereas large q values will promote the redundancy values. Therefore, large q values will show large number of true positives at the cost of introducing additional noise in the Φ, increasing false positives. Therefore, optimal q is the result of a balance between the noise and the attenuation of the redundancy signal and it is obtained using on the cost criteria established, and considering the AUC maximum.

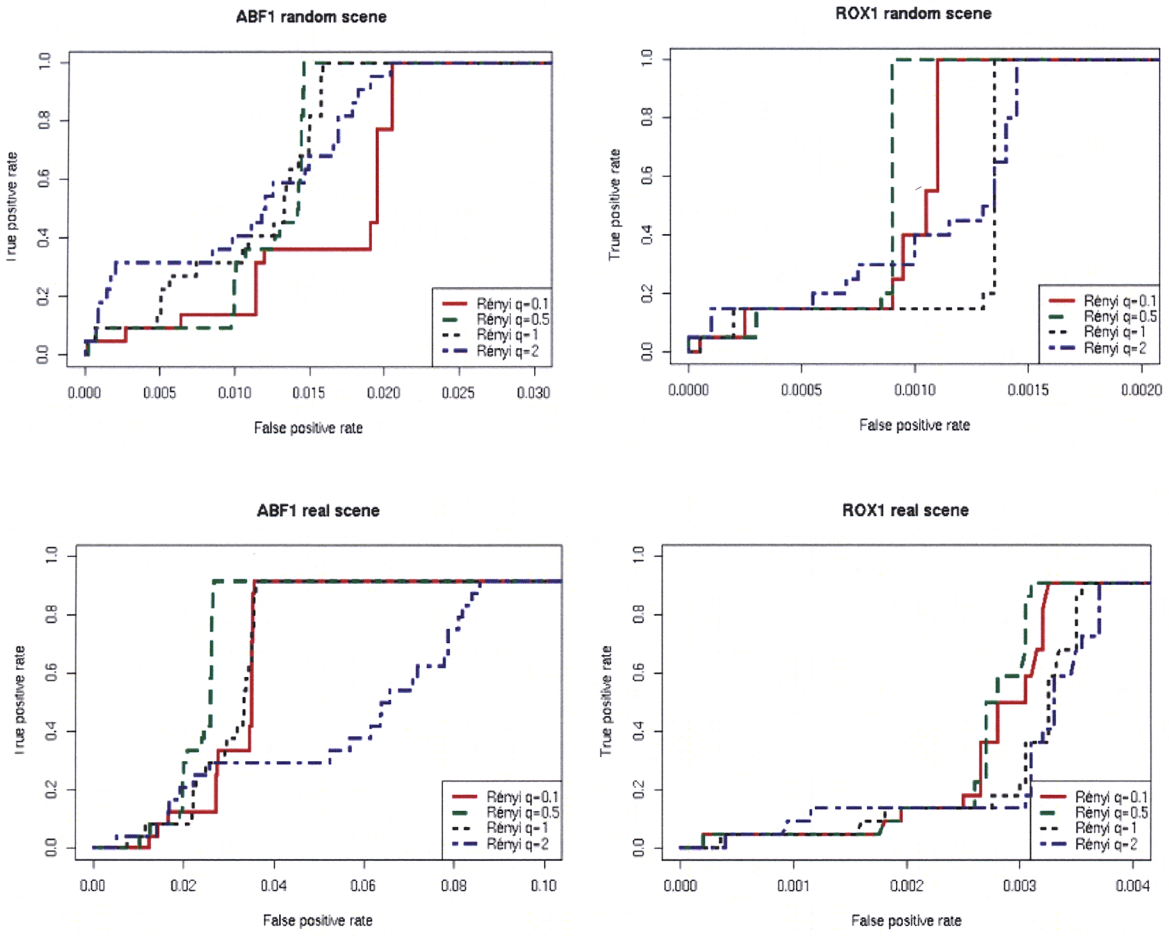

Fig. 2: ROC for ABF1 (left) and ROX1 (right) for random case (up) and *S. cerevisiae* Cr. I and XVI (down) for different q.

IV. Conclusion

In this work, we have presented a method to detect the binding sites of transcription's factor, TFBS, based on a parametric uncertainty measure such as Rényi entropy on a training set sequences corresponding to TFBS. This method has been applied onto two chromosomes of the organism *S. cerevisiae*, seeking binding sites corresponding to ROX1 and ARG1 recognizers. Results suggest that the proposed parametrical uncertainty measurement gives additional information related to binding site detection than Shannon's entropy based detector (q=1). Rényi's order, q, has to be adjusted for every TFBS by means of cross validation. In the process of optimization of the q-value, the redundancy profiles mark the possible positions involved in the binding process.

The detection of binding site is done by the information contained in the training matrix. An incorrect alienation and the lack of information in some positions in the training matrix provokes mistakes in the detection of the binding sites. Moreover, the method used has considered the independence between the positions within the binding site. In future studies, a method of optimal selection of the alienation parameters must be established. It will be also necessary to do a more precise treatment when there's an absence of symbol and it will be possible to consider the dependence between the positions in the binding site by means of base transition frequency using parametric uncertainty measurements [16].

Acknowledgment

CIBER-BBN is an initiative of the Spanish ISCIII.

REFERENCES

[1] H.J. Muller, "The gene material as the initiator and the organizing basis of life". Am. Nat. 100 (1966) 493-517.

[2] R. Mutihac, A. Cicuttin, R. C. Mutihac, " Entropic approach to information coding in DNA molecules", *Materials Science and Engineering*, Vol 18, pp. 51-60, 2001.

[3] R. Mutihac, A. Cicuttin, R. C. Mutihac, " Entropic approach to information coding in DNA molecules", *Materials Science and Engineering*, Vol 18, pp. 51-60, 2001.

[4] T. D. Schneider, "Information Content of individual Genetic Sequences", in *J. Theor. Biol.* Vol. 189, pp. 427-441, 1997.

[5] Geir Sandve, "A survey of motif discovery methods in a integrated framework", *Biology Direct*, 1:11,2006.

[6] Aerts S, Van Loo P, This G, Moreau Y, de Moor B., "Computational detection of cis-regulatory modules", *Bioinformatics* 2003, 19 (Suppl 2): 115-1114.

[7] Sinha S, van Nimwegen E, Siggia ED, "A probabilistic method to detect regulatory modules", *Bioinformatics* 2003, 19 (Suppl 1): i292-301.

[8] Zhou Q, Wong WH, "CisModule: de novo discovery of cis-regulatory modules by hierarchical mixture modelling", *Proc Natl Acad Sci* USA 2004, 101(33):121 14-9.

[9] R. Castelo, R. Guidó, "Splice site identification by idlBNs", *Bioinformatics* 2004, 20 (Suppl I), i69-i76.

[10] Pavesi G, Mauri G, Pesole G., "An algorithm for finding signals of unknown length in DNA sequences". *Bioinformatics* 2001,17(Suppl I): S207-14.

[11] Van Helden J., Rios AF, Collado-Vides J., "Discovering regulatory elements in noncoding sequences by analysis of spaced dyads", *Nucleic Acids Res* 2000, 28(8):1808:8.

[12] Yunlong Liu, Matthew P Vincenti, and Hiroki Yokota, "Principal component analysis for predicting transcription-factor binding motifs from array derived data". *BMC informatics* 2005, 6:276.

[13] Robert A. Gatenby, B. Roy Frieden. "Information Theory in Living Systems, Methods, Applications, and Challenges." Bulletin of Mathematical Biology (2007). 69:635-657.

[14] T.D. Schneider and R.R. Stephens. "Sequence Logos: A New way to Display Consensus Sequences". Nucleic Acid Research, 18:6097-6100, 1990.

[15] Yeo, G. and Burge, C. (2003) "Maximum entropy modelling of short sequence motifs with applications to RNA splicing signals". In Miller, W.(ed), Proceedings of the 7th International Conference on Research in Computational Molecular Biology, ACM Press, NewYork, pp 322-332.

[16] A. Perera, M. Vallverdú, F. Clarià, J. M. Soria and P. Caminal, "DNA Binding Sites Characteritzation by Means of Rényi Entropy Meausures on Nucleotide Transitions". *Accepted for publication on IEEE Transactions on NanoBioSciences*.

[17] Wingender, E. Chen, X. Hehl, R. R. Karas, H. Liebich, I. Matys, V. Meinhardt, T. Pruss, M. Reuter, I. Schachere, and F., "TRANSFAC: an integrated system for gene expression regulation", *Nucleic Acid Res.*, vol. 28, pp.316-319,2000.

[18] Robert C. Edgar, "MUSCLE: Low-complexity multiple sequence alignment with T-Coffee accuracy. 195 Roque Moraes Drive, Mill Valley, CA 94941, U.S.A.

[19] A.Rényi, "A measures of information and entropy", in *Proceedings of the 4th Berkeley Simposium on Mathematics, Statistics and Probability*, 1961, pp. 547-561.

[20] A. Krishnamachari, V. Moy Mandal and Karmeshu, "Study of DNA binding sites using Rényi parametric entropy mesure", in *J. Theor. Biol.*, vol. 227, pp 429-436, 2004.

[21] Debraj Guha Thakurta, "Computational identificacion of transcriptional regulatory elements in DNA sequence". *Nucleic Acids Research*, 2006,. Vol. 34, No. 3585-3598.

[22] T.D.Schneider, G. D. Stormo, L. Gold and A. Ehrenfeuch, "The information Content of Binding Sites on Nucleotide Sequences", in *J. Mol. Biol.* Vol.188, pp.415-431,1986.